基于改进K均值聚类的异常检测算法

左 进 陈泽茂
(海军工程大学信息安全系 武汉 430033)

摘 要 通过改进传统K-means算法的初始聚类中心随机选取过程,提出了一种基于改进K均值聚类的异常检测算法。在选择初始聚类中心时,首先计算所有数据点的紧密性,排除离群点区域,在数据紧密的地方均匀选择K个初始中心,避免了随机性选择容易导致局部最优的缺陷。通过优化选取过程,使得算法在迭代前更加接近真实的聚类类簇中心,减少了迭代次数,提高了聚类质量和异常检测率。实验表明,改进算法在聚类性能和异常检测方面都明显优于原算法。

关键词 K均值,聚类,紧密性,异常检测
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.052

Anomaly Detection Algorithm Based on Improved K-means Clustering

ZUO Jin CHEN Ze-mao
(Information Security Department, Naval University of Engineering, Wuhan 430033, China)

Abstract After optimizing random selection process of the initial cluster centers, an anomaly detection algorithm based on improved K-means clustering was proposed. When the cluster centers are selected, the tightness of all data points is calculated, outliers region is removed, and then the K initial centers in dense regions of data are selected, which avoids that the random selection is easy to cause the defect of local optimum. By optimizing the selection process, the initial cluster centers are more closer to the real clusters centers before iteration of the algorithm, the numbers of iterations are reduced, and the quality of clustering and anomaly detection rate are improved. Experiments show that the improved algorithm is much better than the original algorithm in clustering performance and anomaly detection.

Keywords K-mean, Clustering, Tightness, Anomaly detection

1 引言

异常检测作为数据挖掘领域的一个重要研究方向,主要用来检测数据集中偏离正常分布模式的异常数据。该技术能够从大量的、模糊的复杂数据中提取出潜在的、有价值的信息,在大数据处理中得到广泛应用。现有的异常检测技术主要包含基于监督的和基于无监督的两种方法。基于监督的异常检测方法主要包括概率统计方法、模式预测方法、神经网络方法[1]、增量式SVM异常检测[2];基于无监督的异常检测方法主要包括K-means算法[3]、基于核自适应的AP聚类异常检测算法[4]、引入约束条件的密度聚类异常检测算法[5]。K-means算法作为一种无监督的划分聚类算法,因高效性和简单性被广泛用于异常检测领域。但由于该算法的初始聚类中心选择过程是随机的,因此容易导致最终聚类结果陷入局部最优而非全局最优[6]。为解决初始化问题,Grigorios等人[7]设计了MinMax K均值算法,首先随机选取第一个初始中心,然后在剩下的数据中选取距离第一个初始中心最远的点作为第二个初始中心,其他初始中心的选取都依从以下原则:第i个初始中心是距离前$i-1$个初始中心最小距离中最大的那个。该方法可以使算法的初始聚类中心相互分隔,但不能保证排除离群点。文献[8]提出了类似于细胞分裂的并行二分思想,即使算法自行分裂,直至产生K个初始中心,但该改进方法在分裂过程中具有不确定性,最终选取的初始中心并不是最优中心。朱建宇[9]根据聚类中心的分布规律和收敛目标,提出候选初始聚类中心的概念,扩展了初始中心的选取范围,但额外增加了算法的复杂度。韩最蛟[10]通过计算所有数据点周围的密集性,选择密集性最大的前K个点作为初始聚类中心,但纯粹依据密集性大小选取的初始中心不能保证均匀分布。现有的针对K均值算法初始化过程的改进方法在一定程度上优化了选取过程,但却很少从真实聚类中心的位置和本质属性出发对初始聚类中心进行选取,因而改进之后的算法效果都不太理想。本文从最优聚类中心的性质出发,通过剔除离群点区域,在数据紧密区,对算法的初始聚类中心按照最远距离进行选择,优化了算法的初始化过程,使算法在执行迭代之前获得更合理的初始聚类中心;并基于此,提出了异常检测算法。实验证明,改进后的K-means算法更加高效,基于改进聚类的异常检测算法在异常检测率、误报率和时间复杂度方面也优于原算法。

到稿日期:2015-06-05 返修日期:2015-09-08

左 进(1989—),男,硕士生,主要研究方向为信息安全,E-mail:2944380236@qq.com;陈泽茂(1975—),男,教授,博士生导师,主要研究方向为网络安全。

2 传统 K 均值算法缺陷分析

K-means 算法即 K 均值算法，是 J. B. MacQueen[11] 提出的一种基于划分的聚类方法。该算法由于简单、高效、适用于大规模数据集的处理，被广泛应用于各种领域。K 均值算法属于硬聚类算法，目标是使聚类准则函数(见式(2))收敛。算法结束时，能够将 R^d 空间上的数据集 $X=\{x_1,\cdots,x_i,\cdots,x_n\}$ 划分聚类到 K 个不同类簇中，使得类簇间相似度尽可能小而类簇内相似度尽可能大。

K 均值算法的基本流程是：首先随机指派 K 个数据点作为算法的初始聚类中心即初始类簇中心；然后计算数据集中所有数据点与初始类簇中心的相似度，把各个点划分到与其最相似的中心点所属类簇；对调整后的类簇重新计算其簇中心，再次依据相似度更新所有点的所属簇；如此反复迭代，直至聚类准则函数收敛或达到迭代次数。在该算法中，采用欧氏距离(见式(1))作为数据点之间的相似度。整个算法的时间复杂度为 $O(Knd)$，n 为数据集的大小，K 为聚类簇的个数，d 为数据的维度。

K 均值算法虽然简单高效，但也有缺陷：

(1)初始聚类中心的随机选择使得算法的聚类效果充满不确定性。算法启动迭代时，采用的 K 个初始聚类中心是随机选择的，而没有固定的规则。不同的迭代起点具有不同的搜索路径。因此，聚类的结果对于初始聚类中心有严重的依赖，导致最终的聚类效果容易陷入局部最优而非全局最优。如图 1 所示，若选择的初始聚类中心接近真正的类簇中心，则聚类的结果较为客观真实；如图 2 所示，若随机选取的初始聚类中心包含离群点，则最终的聚类结果会产生较大的误差。

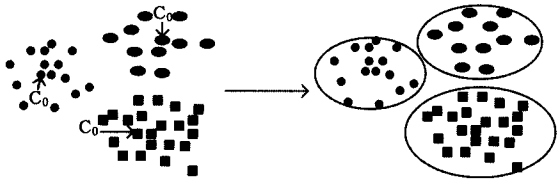

图 1 不同的初始聚类中心对应不同的聚类结果(一)

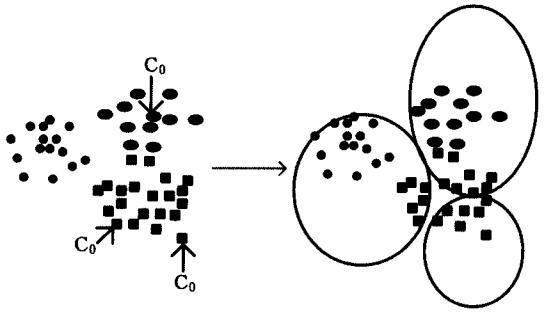

图 2 不同的初始聚类中心对应不同的聚类结果(二)

(2)离群点对聚类结果有重大影响。算法的每次迭代都会根据所有数据点的特征属性进行类簇中心划分，离群点的存在势必会对类簇中心的计算造成很大干扰，影响聚类的结果。

3 基于改进 K 均值聚类的异常检测算法

3.1 对初始聚类中心随机性选择的改进

K 均值算法的第一步是随机选取 K 个数据点作为初始聚类中心，然后启动迭代。对于 K 均值算法，不同的初始聚类中心会导致不同的聚类结果，为了得到更佳的聚类效果，随机性地选取初始聚类中心显然不是好的选择。本文提出的对初始聚类中心选取的改进主要基于以下两个原则：

(1)避免选择离群点。满足这个原则可以使算法一开始就避免陷入误差，使算法产生的结果更加准确。

(2)初始聚类中心在高密度区域选择且分布均匀。显而易见，真实的聚类中心应该是在数据最密集的地方，而且相互保持一定距离。所以，如果初始聚类中心的选择越靠近真实的聚类中心，不仅可以减少迭代的次数，加快收敛，而且可以提高聚类算法的准确性。

为了选取合格的初始聚类中心，改进算法首先计算数据集中所有数据点的紧密程度，剔除较为稀疏的数据区域，得到紧密程度高的数据点集合，因为稀疏区域不仅远离最优的聚类中心，而且还包含离群点；在数据密集区，选择紧密性最高的数据点作为第一个初始聚类中心；然后在该区域内选择距离第一个初始聚类中心最远的数据点作为第二个初始聚类中心；接下来，每一个初始聚类中心即选择为与已选中心最近距离中的最大者，这样可以充分保证各个初始聚类中心的均匀分布。

下面详细介绍初始聚类中心选择的改进算法。

算法 1 初始化过程优化算法

Step1 对于 R^d 空间上的数据集 $X=\{x_1,\cdots,x_i,\cdots,x_n\}$ 中的每一个数据点 x_i，求出其紧密性 $\mathrm{Tigh}(x_i)=\dfrac{1}{\sum\limits_{x_j\in G_t(x_i)} D(x_i,x_j)}$，其中 $G_t(x_i)$ 为 x_i 的 t 个最近邻数据点集合；

Step 2 删除 X 中所有紧密性 $\mathrm{Tigh}<\dfrac{1}{n}\sum\limits_{x\in X}\mathrm{Tigh}(x)$ 的稀疏数据点，得密集数据点集合 X'；

Step 3 在 X' 中，取紧密性最大者即 $\mathrm{Tigh}_{max}(x)$ 的 x 作为第一个初始聚类中心 c_1；取距离 c_1 最远的数据点作为第二个初始聚类中心 c_2；第 $m(3\leqslant m\leqslant K)$ 个初始聚类中心 c_m 为满足以下条件的数据点 $x_i, x_i\in X'$：$\max(D_{min}(x_i,c_1),D_{min}(x_i,c_2),\cdots,D_{min}(x_i,c_{m-1}))$，$i=1,2,\cdots,n$，直至得到最终 K 个初始聚类中心。

通过本文改进方法筛选出的初始聚类中心明显比原始算法随机选择的中心更加具有优势。首先，改进方法排除离群点作为初始中心，保证了算法的迭代起点不会大范围地偏离真实类簇中心；其次，把数据点的紧密性作为初始中心选择的主要依据，符合最优类簇中心的特征；最后，最近最大距离的原理保证了初始聚类中心的均匀分布。

3.2 基于改进 K 均值聚类的异常检测算法

由于 K 均值算法本身的特性，在每次迭代过程中，如果离群点参与聚类簇中心的运算，会给聚类结果带来偏差。因此，本文结合 K 均值算法对离群点敏感这一性质，在改进算法的基础之上提出了异常检测算法。在算法的迭代过程中，检测出异常点，并将其剔除。

算法 2 基于改进 K 均值聚类的异常检测算法

输入：d 维数据集 $X=\{x_1,\cdots,x_i,\cdots,x_n\}$，最终聚类个数 K，聚类函数收敛精度 ε，最近邻个数 t

输出：聚类后的 K 个类簇中心 $C=\{c_1,\cdots,c_j,\cdots,c_K\}$，数据 x_i 所属类簇标签 L，异常点集合 U

Step1 设置初始聚类准则函数值 $J_0=0$，数据集中每个数据点 x 的初始异常度 $Abn_x=0$；

Step2 对于 R^d 空间上的数据集 $X=\{x_1,\cdots,x_i,\cdots,x_n\}$ 中的每一个数据点 x_i，求出其紧密性：

$$Tigh(x_i)=\frac{1}{\sum_{x_j\in G_t(x_i)} D(x_i,x_j)}$$

$G_t(x_i)$ 为 x_i 的 t 个最近邻数据点集合；

Step3 删除 X 中所有紧密性 $Tigh<\frac{1}{n}\sum_{x\in X}Tigh(x)$ 的稀疏数据点，得密集数据点集合 X′；

Step4 在 X′中，取紧密性最大者即 $Tigh_{max}(x)$ 的 x 为第一个初始聚类中心 c_1；取距离 c_1 最远的数据点作为第二个初始聚类中心 c_2；第 m($3\leqslant m\leqslant K$)个初始聚类中心 c_m 为满足以下条件的数据点 $x_i, x_i\in X'$：$\max(D_{min}(x_i,c_1),D_{min}(x_i,c_2),\cdots,D_{min}(x_i,c_{m-1}))$，i=1,2,…,n，直至得到最终 K 个初始聚类中心，分别代表 K 个类簇 w_j，j=1,2,…,K；

Step5 计算 X 中所有数据点与各个聚类中心的欧氏距离：

$$D(x_i,c_j)=\sqrt{\sum_{h=1}^{d}(x_{ih}-c_{jh})^2} \quad (1)$$

其中，i=1,2,…,n 且 j=1,2,…,K。对于数据点 x，若 c_j 使得 $D(x,c_j)=\min D(x,c_j)$，j=1,2,…,K，则将点 x 划分到 c_j 所代表的簇，即 $L_x=w_j$；

Step6 在形成的 K 个类簇中，若属于该簇的数据点 x 与该聚类簇中心距离大于平均距离，即 $D(x,c_j)\geqslant\frac{1}{m_j}\sum_{L_x=w_j}D(x,c_j)$，其中 m_j 是 c_j 代表簇拥有的数据点总数，则 Abn_x++；

Step7 若 $Abn_x\geqslant 3$，则判断 x 为异常点，将其从数据集 X 中剔除，并入集合 U 中；

Step8 判断聚类准则函数

$$J=\sum_{j=1}^{K}\sum_{L_x=w_j}D^2(x-c_j) \quad (2)$$

是否满足收敛条件 $|J'-J|\leqslant\varepsilon$（J 是上次迭代聚类准则函数值，J′是本次聚类准则函数值），若不满足，则转 Step9 继续迭代；若满足收敛条件，则算法结束，输出 C，L 和 U；

Step9 重新计算各类簇的聚类中心：

$$c_j'=\frac{1}{m_j}\sum_{L_x=w_j}x$$

然后转 Step5，m_j 是 c_j 代表簇拥有的数据点总数。

该异常检测算法的基础是 K-means 算法，原理是根据数据点与聚类中心的距离进行异常判断。所以，聚类中心越准确，异常检测性能越好。而 K 均值算法的聚类质量严重依赖初始条件，所以对初始聚类中心的选取尤为重要。通过优化选择过程，选取距离真实类簇中心更近的数据点作为算法起始点，不仅可以提高聚类的质量，而且可以提高异常检测的检测率。

4 仿真实验

实验主要分为两部分，第一部分检验改进后 K-means 算法的聚类性能，主要性能评价指标有：对初始聚类中心选取的合理性、迭代次数和聚类准确率；第二部分分析基于改进聚类的异常检测算法在异常检测方面的性能，主要性能评价指标有：对异常数据的检测率、误报率和算法的平均运行时间。

实验配置：Win 7，VC++6.0，MATLAB7.1，CPU 2.4GHz，内存 2.0GB。实验数据来源于 UCI 机器学习数据库[12]，主要包含 Iris 数据集、Ecoli 数据集、Yeast 数据集，3 个数据集的维数、规模依次增大。实验参数：$\varepsilon=0.25$，$t=10$，测试数据集 Iris，Ecoli，Yeast 时，对应的 $K1=3$，$K2=8$，$K3=10$。

4.1 聚类性能比较

针对 K 均值算法的改进算法众多，如基于模糊聚类的 FCM 算法、K-mean++算法、MinMax K 均值算法等。在本次实验中，为了与 4.2 节以硬聚类算法为基础的异常检测实验保持一致，并且为了能够检验不同算法对于改进初始聚类中心这一步骤的优劣，实验选择了以下 3 种算法作为对比：原 K-mean 算法、MinMax K 均值算法、本文改进算法。

为了验证算法选取的初始聚类中心的合理性，实验采用选取初始聚类中心之后的第一次聚类准则函数值 J_1 来判定。在算法迭代前，若 J_1 越小，说明初始聚类中心越靠近真实聚类中心，选取越合理。同理，迭代次数越少，聚类准确率越高，算法越高效。实验结果如表 1、表 2 所列。

表 1 原算法与改进算法的聚类性能

数据集	K-mean 算法			本文改进算法		
	J_1	迭代次数	聚类准确率(%)	J_1	迭代次数	聚类准确率(%)
Iris	1794.1200	14	78.7	1693.1702	13	86.5
Ecoli	4433.3425	17	76.4	3332.4653	15	84.3
Yeast	8146.7749	26	68.9	4487.4075	20	75.7

表 2 MinMax K 均值算法与改进算法的聚类性能

数据集	MinMax K 均值算法			本文改进算法		
	J_1	迭代次数	聚类准确率(%)	J_1	迭代次数	聚类准确率(%)
Iris	1523.1200	12	79.5	1594.1702	13	87.4
Ecoli	3213.7445	16	76.9	3332.4653	16	85.2
Yeast	9146.7749	28	65.4	5481.4235	21	73.5

实验证明，与原 K-mean 算法相比，由于本文改进算法是通过在数据紧密区域选择初始聚类中心，能够更加靠近真实的类簇中心，因此改进算法的初始聚类准则函数值远小于原算法。而且，初始聚类中心的合理选择也使算法的迭代次数减少，加速了算法的收敛。最终，改进算法的聚类准确率也远远高于原算法。

与 MinMax K 均值算法相比，由于 MinMax K 均值算法侧重于初始聚类中心的均匀分布，因此该算法能够迅速达到局部最优解，第一次聚类准则函数值 J_1 和迭代次数方面都优于本文改进算法，但聚类准确率不如本文改进算法；并且在处理高维和大规模数据集时，本文改进算法的性能明显优于 MinMax K 均值算法。

4.2 异常检测分析

这一部分实验主要测试算法对数据集异常检测的检测率、误报率和平均运行时间。通过在 3 种不同维数和规模的数据集中人工加入一定比率的异常数据，对算法进行测试。实验结果如表 3、表 4 所列。

表 3 原算法与改进算法的异常检测性能

数据集	基于 K-mean 异常检测			基于改进 K-mean 异常检测		
	检测率(%)	误检率(%)	平均用时(ms)	检测率(%)	误检率(%)	平均用时(ms)
Iris	75.3	16.8	999	83.4	5.4	936
Ecoli	70.2	23.5	1026	78.3	9.6	975
Yeast	66.0	31.4	1786	72.5	10.2	1232

表 4　MinMax K 均值算法与改进算法的异常检测性能

数据集	MinMax K 均值算法异常检测			改进 K-mean 异常检测		
	检测率 (%)	误检率 (%)	平均用时 (ms)	检测率 (%)	误检率 (%)	平均用时 (ms)
Iris	70.3	20.8	892	85.6	6.3	916
Ecoli	68.1	25.6	946	79.3	8.9	1025
Yeast	55.2	36.5	1566	73.4	11.2	1332

实验表明，与原 K-mean 算法相比，无论是在检测率还是误报率方面，改进算法都优于原算法；而且在算法的平均运行时间方面，改进后的算法也更加高效。由于原算法选择聚类中心的随机性，导致算法可能选择异常点或其附近的点作为初始聚类中心，从而使聚类结果产生较大误差，因此改进算法在误检率方面明显优于原算法。

与 MinMax K 均值算法相比，改进算法用时偏多。但 MinMax K 均值算法运行得到的结果并非最优结果，所以在检测率和误检率方面，该算法性能低于改进算法。

结束语　本文通过均匀选择数据紧密区域及避免离群点区域，优化了 K 均值算法的初始聚类中心选取过程，提出了基于改进 K 均值聚类的异常检测算法。实验证明，改进之后的算法不仅在聚类性能方面更加高效，而且在异常检测方面也比其他算法更加具有优势。随着越来越多大数据的产生，研究面向更高维度、更大规模、适合数据流形式的数据挖掘算法将是本文下一步努力的方向。

参 考 文 献

[1] Yang Yu-zhou. Research and implementation of the clustering anomaly detection technology based on feature extraction[D]. Chengdu: University of Electronic Science and Technology of China, 2012(in Chinese)
杨宇舟. 基于特征提取的聚类异常检测技术的研究与实现[D]. 成都：电子科技大学，2012

[2] Sun Na, Guo Yan-feng, Yao Yuan. Network data stream abnormal detection model based on SVM incremental learning method [J]. Computer Engineering and Applications, 2012, 48(29): 78-81(in Chinese)
孙娜，郭延锋，姚远. 增量式 SVM 的数据流异常检测模型[J]. 计算机工程与应用，2012，48(29)：78-81

[3] Luo Yong-jian. Research on Data Flow Anomaly Detection Algorithm Cluster-based[D]. Harbin: Harbin Engineering University, 2010(in Chinese)
骆永健. 基于聚类的数据流异常检测算法的研究[D]. 哈尔滨：哈尔滨工程大学，2010

[4] Fu Ying-ding, Lan Ju-long. Kernel-based adaptation for affinity propagation clustering algorithm[J]. Application Research of Computers, 2012, 29(5): 1644-1650(in Chinese)
付迎丁，兰巨龙. 基于核自适应的近邻传播聚类算法[J]. 计算机应用研究，2012，29(5)：1644-1650

[5] Jiang Min, Pi De-chang, Sun Lan. Research on Density Clustering Algorithm with a Multiple Constraints[J]. Computer Science, 2011, 38(10A): 143-164(in Chinese)
江敏，皮德常，孙兰. 一种多约束的密度聚类算法的研究[J]. 计算机科学，2011，38(10A)：143-164

[6] Celeb M, Kingravi H, Vela P. A Comparative Study of Efficient Initialization Methods for the K-methods for the K-Means Clustering Algorithm [J]. Expert Systems with Applications, 2013, 40(1): 200-210

[7] Tzortzis G, Likas A. The minmax k-means clustering algorithm [J]. Pattern Recognition, 2011, 44(4): 866-876

[8] Jiang Da-yu. A fast and efficient parallel bisecting K-Means algorithm[D]. Harbin: Harbin Engineering University, 2013 (in Chinese)
蒋大宇. 快速有效的并行二分 K 均值算法[D]. 哈尔滨：哈尔滨工程大学，2013

[9] Zhu Jian-yu. Research and Application of K-means algorithm [D]. Dalian: Dalian University of Technology, 2013(in Chinese)
朱建宇. K 均值算法研究及其应用[D]. 大连：大连理工大学，2013

[10] Han Zui-jiao. An Adaptive K-means initialization method based on data density[J]. Computer Applications and Software, 2014, 31(2): 182-187(in Chinese)
韩最蛟. 基于数据密集性的自适应 K 均值初始化方法[J]. 计算机应用与软件，2014，31(2)：182-187

[11] Macqueen J. Some methods for classification and analysis of multivariate observe[C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeyey: University of California Press, 1967: 281-297

[12] Asuncion A, Newman D. UCI Machine Learning Respository [EB/OL]. [2015-06-01]. http://archive.ics.uci.edu/ml/datasets.html

(上接第 232 页)

[4] Zhu Y X, Zhang X G, Sun G Q, et al. Influence of Reciprocal Links in Social Networks[J]. PLoS ONE, 2014, 9(7): e103007

[5] Cha M, Haddadi H, Benevenuto F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy [C]// International AAAI Conference on Weblogs and Social Media(ICWSM). 2010

[6] Romero D M, Galuba W, Asur S, et al. Influence and Passivity in Social Media [C]// ECML PKDD. 2011

[7] Gong Shang-fu, Chen Wan-lu, Jia Peng-tao. Research on clustering algorithm of hierarchical clustering[J]. Computing Applications, 2013, 30(11): 3217-3218(in Chinese)
龚尚福，陈婉璐，贾澎涛. 层次聚类社区发现算法的研究[J]. 计算应用研究，2013，30(11)：3217-3218

[8] Duan Ming-xiu. Research and application of hierarchical clustering algorithm[D]. Changsha: Central South University, 2009(in Chinese)
段明秀. 层次聚类算法的研究及应用[D]. 长沙：中南大学，2009

[9] Zhang Li-hua. Application of data mining technology in special celestial discovery[D]. Jinan: Shandong University, 2009(in Chinese)
张丽华. 数据挖掘技术在特殊天体发现中的应用研究[D]. 济南：山东大学，2009

[10] Hou Bin. Typical correlation analysis algorithm based on sparse representation[D]. Nanjing: Nanjing University of Science and Technology, 2013(in Chinese)
侯彬. 基于稀疏表示的典型相关分析算法研究[D]. 南京：南京理工大学，2013